



下载APP



01 | 平均值：不要被骗了，它不能代表整体水平

2021-07-28 郭炜

《数据分析思维课》

课程介绍 >



讲述：郭炜

时长 17:25 大小 15.96M



数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

我们在日常生活中经常会遇到这种情况，看到某些统计报告里面说，“某市的人均住房面积是 120 平米”“计算机行业人均年收入超过 50 万元”。你看完这个数据之后，倒吸一口凉气，然后去微博感叹：“对不起，我又给大家拖后腿了”“对不起，我又被幸福了”。

你不必为此焦虑，我只能奉劝你，以后看到这么不专业的统计报告就别看了。来，我带你看看准确客观的平均值统计应该是怎样的。首先，你得知道平均值究竟是什么。



从概念上看，平均值有很多种。单从数学上来说，就有算术平均值、几何平均值、平方平均值、调和平均值、加权平均值等等。所以当有人和你谈平均值的时候，你要留个心眼问

问他，你这说的是哪个平均值呀？

当然，我们日常生活中提到的平均值都默认是“**算术平均值**”，也就是“**一组数据中所有数据之和再除以数据的个数**”。这个概念不难理解，你在小学的时候就开始学了。不过看到这里，你可以先结合我们上面的例子想想，算术平均值有什么短板吗？

我先给一道极其简单的数学题，你可以先想想。我们有 3 个数，他们分别是 0，1，20，这三个数的平均值不难算，是 $(0+1+20)/3=7$ ，那 7 这个平均值和之前的三个数是不是差距挺大呢？是不是有些不客观呢？

所以，有的时候，平均值并不能代表整体水平。

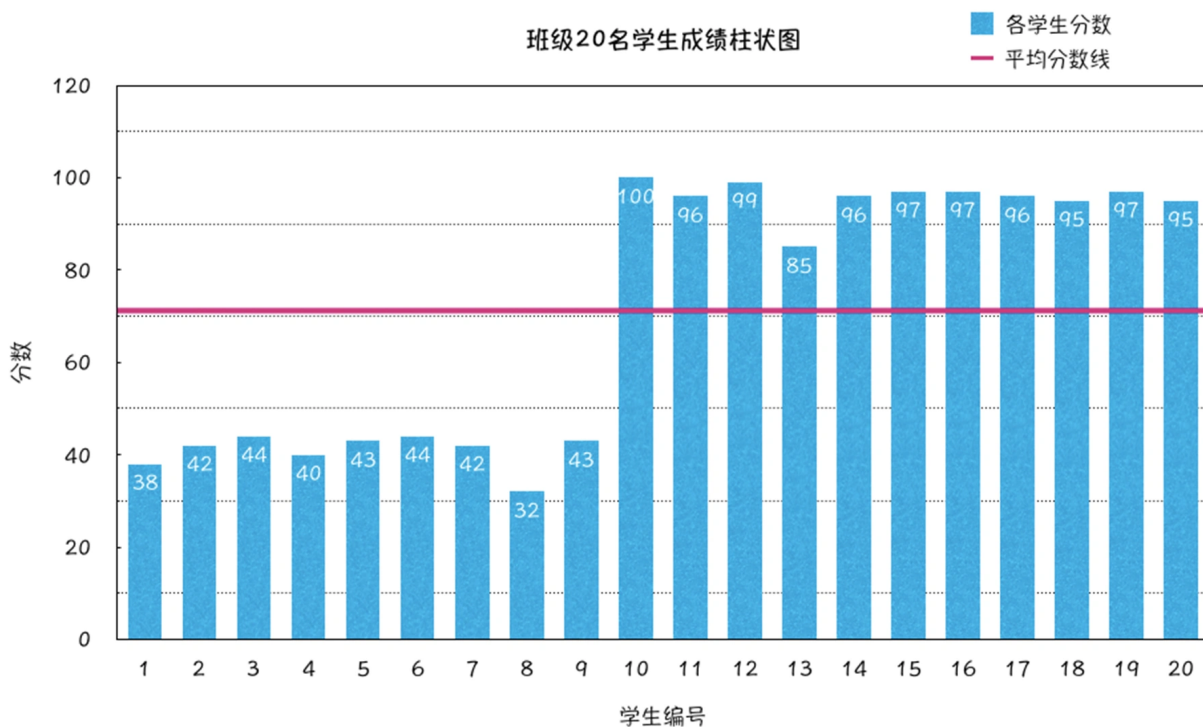
平均值在什么情况下才有价值？

那平均值到底在什么情况下才有价值呢？回答这个问题之前，我再给你讲个故事。

昨天下楼的时候，我听到小区两个大妈在讨论，“这次期末考试，班里语文的平均分是 71 分，我孙子考了 85 分，厉不厉害！”在工作中，我偶尔也会听到同事说，“我们客户的平均客单是 1000 元钱，竞争对手的只有 500 元，我们的客户比对方的高端多了。”这些说法都对吗？还真不一定。

为了更好地解释这个问题，我先拿孩子的平均分给你举个例子。假设班级里 20 名学生的考试成绩如下图一样呈现两极分化的情况，一半孩子都在 95 分以上，还有近一半的孩子只有三四十分，我们很容易计算出这 20 名学生的成绩平均值是 71.05（图中的红色直线）。

看上去孩子的 85 分比平均分 71.05 分高了很多，但你再仔细看，这个分数在好学生里其实是最差的那个，整体上看也只是班级中游水平。



同理，看上去这个企业的平均客单价很高——平均 1000 元，但如果你的数据是由 1 个 1 万元客户和 10 个 100 元的客户构成的（总收入 11000 元 / 11 人 = 1000 元 / 人），对方都是 11 个客单价 500 元的客户构成的，那么其实竞争对手才是真正的高客单价企业。

你可能会觉得我在抬杠，这些例子的数据集都太极端了吧？其实我是想说明一个问题：平均值是用所有样本数据计算的，**容易受到极端值的影响**。在不少情况下，平均值是没有价值的，它无法客观准确地反映数据整体情况。

更进一步来说，**整体平均值是在数据呈均匀分布或者正态分布的情况下才会有意义，如果忽略整个数据的分布情况，只提平均值，其实是没有意义的**。这也就是为什么你会在读一些统计分析报告时觉得自己不是“被加薪了”，就是“被幸福了”。

现在你明白了吗？在一些复杂情况下，我们是很难确定人群分布情况的，此时若直接使用平均数值，是很难反映整体真实情况的。

分组结论和整体平均值不是一回事

那怎么才能反映真实情况呢？

就拿平均薪水这个例子来说，你肯定有疑问：什么人啊？咋拿到那么多钱的？你肯定想看更详细的数据，诸如具体的岗位属性、工作年限、城市等等。有了这些信息，你才能知道你和人家的薪水差距到底差在哪了。

比如一个在一线城市工作 3 年的 Java 程序员的月平均工资是 2 万元，而我的月工资是 1 万元，那我确实是低一些，这个判断比起之前那个一刀切就准确太多了。

不知道你有没有注意到，在思考这个问题的过程中，你已经在不知不觉中引入了**分组**的逻辑。你应该也发现了，分组中的平均数和从整体中得到的平均数，是完全不一样的，分组中得到的平均数更具参考价值。

上面这个例子很好理解，我现在要顺着它抛出一个结论：**整体平均值不能代表各分组情况，分组结论和整体平均值结论可能会大相径庭。**

明白了吗？别急，我再讲一个例子反面论证一下这句话。话说 NBA 有两个球员，球员 A 和球员 B，他们的投球的表现如下面这个图所示。

NBA两位球员投球表现				
	球员A		球员B	
	2分球	3分球	2分球	3分球
投中	200	5	90	50
总投球数	250	50	100	150
投中率	80.00%	10.00%	90.00%	33.33%



这里我简单描述下，先说 2 分球的情况：A 球员，2 分球总共投了 250 个，投中了 200 个，命中率 80%；B 球员，投了 100 个，投中了 90 个，命中率 90%。也就是说，以 2 分球的命中率来看，B 更牛一些。

投 3 分球的时候，A 球员一共投了 50 个，投中 5 个，命中率 10%；B 球员，一共投了 150 个，投中 50 个，命中率 33.33%。看来，无论 2 分球还是 3 分球，B 都比 A 的投中率要高。看上去也是 B 比 A 厉害，对吧？

那问题就来了，可是从整体命中率来看好像不是这样啊。你看，如果我们算下两位球员的整体平均值（也就是整体投中率）。

NBA两位球员整体投中率		
	球员A	球员B
投中	205	140
总投球数	300	250
投中率	68.33%	56.00%

 极客时间

A 的总投球数是 300 个（250 个 2 分球，50 个 3 分球），共投中 205 个（200 个 2 分球，5 个 3 分球），投中率是 68.33%；B 呢，投了 250 个球，投中的两分和三分加到一起 140 个，那么他的投中率是 56%，这么看来 B 比 A 的投中率 68.33% 要低呀。

看到这个结果，你是不是很诧异，不信你再仔细看看这个图表，我的确没有在数据上做手脚。

两分球和三分球投中率都比较高的这个球员 B，整体的命中率反而下降了，是不是让人有点大跌眼镜？如果你是篮球爱好者，你应该会发现问题，NBA 里没算过整体命中率，一般都是把二分之一球和三分球的命中率分开说的。

辛普森悖论的启示

为什么 NBA 不计算整体命中率呢？就是因为这样算不准确。这里我们可以引入一个著名的悖论——**辛普森悖论**，它讲的就是这个问题。辛普森悖论是 1951 年由 E.H. 辛普森提出的，简单来讲就是在**分组比较中都占优势的一方，有的时候在总评中反而是失势的一方**。

我用前面 NBA 球员命中率的例子跟你分析下产生辛普森悖论的原因。首先，两分球和三分球的投球能力根本不是一回事，这两个投球数本身就不应该相加；另外，B 球员虽然厉害，但是他 60% 的球都是命中率比较低的三分球，而绝对数量上命中率较高的两分球投得就少了，而三分球的投中率是明显低于两分球的，这样就拉低了他的整体的投中率，造成了整体的劣势。

是不是还是有点绕？用一句话来讲，就是“质”（命中率）与“量”（投球数）是两个维度的数据，如果全部合并成“质”（命中率）这个维度的数据，那就会出错了。

再举一个例子，某游戏公司做了款游戏，有 Android 和 iOS 版本，而每个系统都有手机版本和 Pad 版本。一个数据分析师看完用户的付费数据后，发现整体上 Android 付费率比较高。他就直接告诉老板说，“你看我们 Android 的用户付费率要高于 iOS 的用户付费率，我们应该大力发展 Android 客户端！”这个数字是真实的，但是结论很可能是错误的。

因为我们再细分下去你会发现，这个数据分析师也是错误地把“质”（付费率）和“量”（用户数）简单合并，是一种想当然的行为。

因为还有可能出现这样一种情况：Android 无论是平板还是手机的付费率分别都比 iOS 低，但是整体上因为安卓手机（注意，只是手机用户）的用户比较多，所以把 Android 付费率整体拉高了。但其实细分下去 iOS、Pad 和手机的付费率都比 Android 高，只是整体付费率低而已。你要是还有疑虑，可以对照上面 NBA 的例子，自己再推演一下。

所以，我再来总结下，看到一个平均值的时候，你一定要留个心眼，看看它的数据构成情况，而不是简单地用平均值去代表所有的整体。生活是具体的，如果你想看到更为准确的数据，你应该分组拆开来看。因为辛普森悖论告诉我们，有的时候，在分组比较中占优势的一方，在总评中反而可能是失势的一方。但你要注意，只是“有的时候”。

就像最近我看到一些文章，说税率改革之后我们的整体工资的税率反而是变高了，而不是变低了。这也是同一个道理，我们用整体的平均值去掩盖个体每一个不同区间段的税率变化，是不对的。我们应该更细分的数据去评定实际的好坏。

除此之外，辛普森悖论也给我们一个启示，就是：**每次小范围内的输赢，其实和你在整体上的输赢没有太大直接的关系。**这也是为什么在打麻将或者打德州扑克真正赢的那些人，不是

那些小牌把把赢的人，而往往是赢一把大的人。

这也是这个辛普森悖论衍生出来的一个推论，将来你要用数据分析做决策的时候，小到打牌、大到做投资，不要过于计较局部的得失，而是要在关键时刻对大概率有把握的事情放手一搏。

小结

好了，今天这篇文章就到这里。我在最后再来给你串讲下这节课的知识点。

首先，当别人给你说平均值的时候，你要和他确认下说的是哪个平均值。当然，生活中，我们提到的平均值基本都是在说**算术平均值**。其次，算术平均值特别敏感，它很容易受到**极端数据**的影响，所以在很多选秀节目里，你经常会听到最后计算分数时要去掉一个最高分和一个最低分，这是一个道理。

你也一定要意识到，整体平均值是在**数据均匀分布或者正态分布**下才会有意义，如果忽略整个数据的分布情况，只提平均值是没有价值的。

最后，我和你聊了辛普森悖论。工作生活中，我们经常会遇到这样的悖论，甚至我见过很多传销人员就在用这个悖论在忽悠人，如果你遇到这样的案例，别忘了那句话：**分组结论和整体平均值结论可能会大相径庭**。

在我们的生活里，我们总提“质量”这个词，但是拆开来看，“质”与“量”是不等价的。所以当你不被大部分人所理解时，有可能是因为你选的路是一条少数人走的路。平均值和辛普森悖论告诉我们要抓大放小，不要因为某一个单项优势就洋洋得意，也不要因为局部失败就一蹶不振。生活，要有一颗平常心，我们的目标是让我们这一生的“人生平均值”逐步提高。

数据给你一双看透本质的双眼，让我们持续学习，持续提高。

课后思考

最后我给你留一个课后思考题：你在你的生活里，你还遇到过哪些平均值和辛普森悖论的例子吗？欢迎你分享出来，我们一块讨论。

欢迎在留言区与我分享你的想法，也欢迎你在留言区记录你的思考过程，如果你能有其他案例，那就更好了。感谢阅读，如果你觉得这篇文章对你有帮助的话，也欢迎把它分享给更多的朋友。

分享给需要的人，Ta订阅后你可得 **20** 元现金奖励

👍 赞 11 💡 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 导读 | 万物背后的规律都是数据

下一篇 02 | 大数定律与小数陷阱：生活是随机还是有定数的？

精选留言 (13)

💬 写留言



Juha

2021-07-29

哈哈，重新读一下，很real

展开 ▾

💬 2

👍 3



mtfelix

2021-08-01

大侠的课，对活法也极有启发呀。赢一堆鸡毛蒜皮，不如一把大的。

作者回复: 数据思维，其实思维启发更是一大部分，知识是死的，思维是活的，要听我亲自录的音更有感觉，哈哈



💬

👍



mtfelix

2021-08-01

2，3分球那个例子太棒了，同理，假设生活中真牛人更多做挑战性大的事，那肯定总成功率并不一定高，但确实是牛。

**进化菌**

2021-08-01

在生活里，其实大部分时候说的就是算术平均值，这可能跟其他平均值了解的人不多有关吧。

那么，我们既然知道算术平均值的缺陷，而且平均值有更多的考虑维度，平时就没必要过于叹息了。对比的价值在于认识自我，而不是进一步的打压自己。

...

展开

作者回复: 是的，不能看局部，要看整体。

**夜路破晓**

2021-07-31

之前只是从量的角度理解辛普森悖论，比如生活当中存在一类聪明人，在整体水平较低的模块占据优势，因此获得重视，但对基础模块的提升只是浅尝辄止，那么很有可能导致个人整体水平不高或者存在上限，因而走不远。

看完这篇文章对辛普森悖论的本质有了更进一步的了解。人们在认识世界方面，由于客观条件的限制必然存在着多种视角，沟通交流就是对建立在对多种角度融合的基础上。那...

展开

作者回复: 赞同！后面讲聚类 and 关联算法背后的哲学逻辑也会提到，断舍离，人的一生非常短暂，在值得投入的人和事情上面深度投入，不值得的事情一分钟都嫌多。



1

**Troy@InfoQ_0a1dfd51...**

2021-07-31

单从投篮命中率会不会也不够，假如A跟B上场时间差不多，A的总得分是415而B是330，这样看A是更有得分能力

作者回复: 你说的对！数据分析是最后是用结果说话的，进球时间效率以及其它队友进球和防守配合情况要通盘考虑，而且要看整体球队，而不是某个明星球员，要看整个赛季而不是某场比赛，这就是辛普森悖论教给我们的。



**小呀么小二郎**

2021-07-31

- 1、平均值只有在数据均匀分布或正态分布才有意义。
- 2、辛普森悖论。
- 3、分清质与量。【感觉这个比较难】
- 4、把握整体，交给时间。

**公号-技术夜未眠**

2021-07-30

赢得了每一场战役，却最终输掉了整个战争。

展开 ∨

作者回复: 是的，要有格局



👍 1

**第一装甲集群司令克莱...**

2021-07-30

先胖不算胖，后胖压倒炕。
打麻将开局顺风顺水，一堆屁胡。
然后人家几把大胡，是最后赢家！

展开 ∨

作者回复: 哈哈，先胖趋势不能代表整体水平。

**mapping**

2021-07-29

懂了，这就去梭哈 😊

展开 ∨

**Geek_fbe1ea**

2021-07-29

平均值-均匀/正态/极端，分组-整体，辛普森悖论。原来以为已经懂了，引入分组与平均值，对辛普森悖论有了更深的理解

展开 ∨

作者回复: 是的，其实过去我们学的东西都在书本上，学校却没有教我们怎么真正运用在生活和工作里，数据博大精深，一起学习，如果觉得好，也欢迎分享给他人



吴婧雯 J
2021-07-29

小白诚心发问。请问在数据分析的学习和职场路径上，是否有除了学位类的证书/认证(范围大较大和更普及类的)？感谢！

展开

作者回复: 国内数据分析的认知还刚刚起步，其实很多知识都是我们学校里学过的，知识领域来讲，这门课覆盖的内容，如果深入学习已经足够了，关键是要学以致用，在工作和生活中用起来，这才是数据分析的真谛



一步
2021-07-28

量变引起的质变吗？

展开

作者回复: 另外一节课会讨论，量和质的关系，